Classification of Fuzzy Data in Database Management System

Deval Popat[1], Hema Sharda[1], and David Taniar[2]

[1]School of Electrical and Computer Engineering,
RMIT University, Melbourne, Australia
Phone: +61 3 99255360, Fax: +61 3 99255340
s2011159@student.rmit.edu.au
hema.sharda@rmit.edu.au
[2]School of Business Systems,
Monash University, Melbourne, Australia
David.Taniar@infotech.monash.edu.au

Abstract. When the available information is imperfect, it is often desirable to represent it in the database, so that it can be used to answer queries of interest as much as possible. The data as well as query in data sources are often vague or imprecise (fuzzy). In this paper, a comprehensive classification of fuzzy data is done. This classification will be used as framework for understanding how fuzzy data arise and manifest themselves.

1 Introduction

Computers do not reason the way human minds do. They simply manipulate solid facts into ones and zeros whereas the human mind can process incomplete data with uncertainties and unknowns and still function perfectly. Thus came the concept of Fuzzy Logic in order to create machines that can imitate the complex functions of the mind.

1.1 Fuzzy Database Evaluation

Fuzzy logic provides the solution by providing a mathematical approach to programming complex systems [1]. It implements operator knowledge directly into the system. The thought process of how humans store and apply knowledge about controlling a system can take the form of "if <situation> then <reaction>" rules. For example, a process may require a rule such as "if the temperature is rising fast than stop heater". Since the rule is quite abstract this would be difficult to program into a traditional system.

Fuzzy Logic is a method for grouping items into sets, which does not require that an item be either in or out of a set completely [5]. Fuzzy sets can be characterised by a *fuzzy membership function*, a function which takes a number of attributes of the items

M.Gh. Negoita et al. (Eds.): KES 2004, LNAI 3214, pp. 691–697, 2004.

to be classified as input, and provides a fuzzy membership value between 0 and 1 as output.

The need for fuzziness in data modelling can be viewed as providing the capability to directly represent imprecision. Clearly now that AI applications are becoming more common, the utilization of database technology becomes more critical if these applications are to reach their full potential and usefulness. Below we summarise some of the reasons of using fuzzy logic in Database:

Fuzzy logic provides a user-friendly data presentation for the report generation phase with linguistic variables and fuzzy values.

Fuzzy Sets provides additional data security features due to the introduction of an additional view-based data layer, which hides the numerical values from the users. At the same time, the information is available for querying or evaluating the very large databases by using linguistic variables.

Fuzzy logic provides optimised database performance by modelling imprecise, uncertain and vague data.

2 Classification of Fuzzy Data

Many real world systems and applications require information management components that provide support for managing imprecise data. Fuzzy theory allows us to model imprecise or vague data. The use of fuzzy theory also allows us to model vague knowledge. There have thus been several proposals for extending relational database systems in order to represent as well as query such imprecise data. Little work, however, has been done in modeling uncertainty at the conceptual schema level and in standardizing fuzzy data in fuzzy relational databases (FRDBs). To fill this gap, a classification of fuzzy data is developed. This methodology contains extensions for representing the imprecision of data in fuzzy relational database. We limit the scope of the paper with the following assumption that the types of data considered in this paper are only numerical and string data. In particular, we exclude multimedia data from consideration.

For many practical systems, important information comes from two sources: one source is human experts who describe their knowledge about the system in natural languages; the other is sensory measurements and mathematical models that are derived according to physical laws. Doctors, lawyers, engineers can diagnose problems a lot quicker if the expert system they use to diagnose the problem lists a few fuzzy solutions that they can use to augment their own findings. To store expert opinions, fuzzy database is necessary which stores fuzzy data (linguistic terms).

Fuzzy data means imprecise, vague data or non-standard representation of the same data. Imprecise, vague, uncertain, ambiguous, inconsistent, incomplete and null data are fuzzy data. Here we discuss each type of fuzzy data in detail.

2.1 Imprecise Data

Imprecise data is not erroneous and does not compromise the integrity of an information system. Imprecision arises from the existence of a value, which cannot be measured with suitable precision. Specific kinds of imprecise information include:

2.1.1 Disjunctive Data

Disjunctive data means data that presents two or more alternative terms. If the theory DB has several distinct models M_1, M_2, …., M_n, we are not certain whether W is M_1, or M_2 or …. or M_n. This can be expressed by $W = M_1$ or $W = M_2$ or …. or $W = M_n$ [9].

For example, in Table 1, to reach ST Kilda we can catch either tram 16 or 96. In fuzzy database both value has same possibility for reaching the destination so degree of membership will be one for both trams. Fig. 1 represents Disjunctive Data.

2.1.2 Negative Data

Negative data means denying or contradicting data. Negative data implies range of possible values as it denies one value but the possible value can be other than that. A subset W of a set U can be represented by a function X_W: U → {0,1} than negative information can be presented as $X_W'(x) = 1 - X_W(x)$. As shown in Table 2, the fuzzy set will have membership value of 0 for negative data but membership value will vary for the range of all other possible values. Fig. 2 represents Negative Data.

Table 1. Route Table

Destination	Tram No
ST Kilda	16 or 96
Thornbury	112 or 86

Table 2. Population Table

Family Name	No of children	No of boys
Patel	2	None
Gupta	2	1

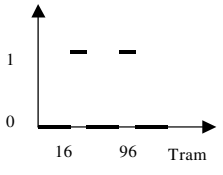

Fig. 1. Disjunctive Data

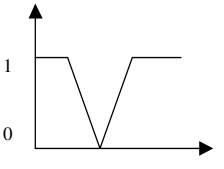

Fig. 2. Negative data

2.1.3 Range Data

Range data means data that vary within or over specified limits. If the theory DB has several distinct models M_1, M_2, …., M_n, and we know that W is equal to one of the values from M_1 to M_n. This can be expressed by $W \in M_1, M_2 \ldots$ to… M_n. For example in Table 3, Fuzzy set of data will be developed based on membership function for given range. Membership value will be assigned to each data value. Fig. 3 represents Range Data.

2.1.4 Data with Error Margin

Data with error margin means a limit in a condition, beyond or below which data is no longer possible or acceptable. If the theory DB has several distinct models

M_1, M_2, …., M_n, and we know that W is equal to one of the value from M_1 to M_n and its variation δ. This can be expressed by $W \in M_i \pm \delta$ where i = 1,…..,n.

For example, Table 4 gives information about which instrument allows how much error margin. In Fuzzy database, upper limit and lower limit of the fuzzy set is set by given margin of data. Membership function will be written on the basis of given margin and membership value will be assigned to each data value. Fig. 4 represents Data with Error Margin.

Table 3. Student Table

Student No	Age
2011159S	25-32
2121157T	18-25

Table 4. Instrument Table

Instrument Type	Measure-ment	Allowed Error Margin
Weighing machine	70kg	0.2kg

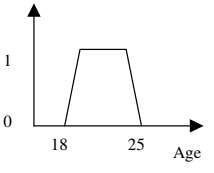

Fig. 3. Range Data

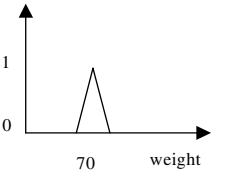

Fig. 4. Data with Error Margin

2.1.5 Null Data

The two extreme kinds of imprecision are precise values and null values: a value is precise when the set of possibilities is a singleton; a *null value* usually denotes that no information is available, yet could be regarded as imprecise information where the set of possible values encompasses the entire domain of legal values. A basic problem with null values is that they have many plausible interpretations. Most authors agree that the various expressions of nulls can be reduced to two basic interpretations. These are:

The unknown interpretation: a value exists but it is not known. Additional information may be available on a specific instance of a null [10]. *The nonexistent interpretation*: a value does not exist. A value is undefined. This meaning of null, however, is not related to uncertainty or fuzziness.

2.2 Vague Data

Vague means not clearly expressed or not clear in meaning or application. Vague data contains some vague predicate such as “tall”. When modelling the concept “tall” as a fuzzy subset of [0,∞) with a membership function A: [0,∞) → [0,1], is a description of the meaning “tall” in a mathematical way.

For example, "Tom is tall". The statement could be used to establish the range of possible values for Tom's height, based on fuzzy set interpretation of term "tall". Different Individual can interpret the word "tall" differently.

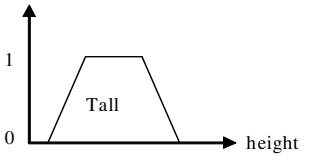

Fig. 5. Vague Data. Fuzzy set interpretation of term "Tall"

2.3 Uncertain Data

Uncertainty arises from the fact that an agent has constructed a subjective opinion about the truth of a fact, which it does not know for certain. This lack of information makes it impossible to determine if certain statements about the world are true or false, all that can be done is to estimate the tendency of the statement to be true or false by using some numerical measure of the degree to which one may be sure [6].

2.3.1 Uncertain Data Due to Statistical Analysis

Some data is recorded statistically and so is inherently uncertain [4].

2.3.2 Uncertain Data Due to Security Reasons

Other data is deliberately made uncertain for security reasons [7,8]. Other data may not be measured accurately, due to some quantum mechanical effect, and will include some irreducible uncertainty. In such situations, the best that we can do is to try to estimate the tendency of the statement to be true (or to be false). This can be done with the help of fuzzy set & by providing degree of membership to the statement to be true or false.

2.4 Ambiguous Data

Ambiguous means doubtful, uncertain, or capable of being understood in either of two or more possible senses.

Table 5. Building Table

Type of room	Length	Width
DR	40 meters	20 meters
DR	10 meters	10 meters

2.4.1 Ambiguous Data Due to Use of Abbreviation

For example, Building Table. Here confusion is that DR should be interpreted as Dinning Room or DR should be interpreted as Drawing Room.

2.4.2 Ambiguous Data Due to Incomplete Context

For example, Weather Report Database. It might show that today temperature will be 30º. Here confusion is whether temperature is in °C (Celsius) or °F (Fahrenheit)?

2.4.3 Ambiguous Data Due to Different Orderings

Ambiguous data occur when different values for the same item are stored by the system. For example, Bill Clinton vs. Clinton, Bill.

2.5 Inconsistent Data

Inconsistent data means, data that doesn't agree with itself or which is not reliable or data that does not have one common solution, as of two or more equations. For example, in the context of distributed databases, if each database is considered an agent, it may happen that A is derivable from the database DB_1 and $\neg A$ is derivable from the database DB_2. In this case uncertainty is about the database that contains the correct information. The mutual inconsistency of DB_1 and DB_2, that is, their lack of common model, can be expressed by $W \in M(DB_1)$ or $W \in M(DB_2)$.

2.5.1 Inconsistent Data Due to Unreliable Sources

These types of data exist because the sources of data are unreliable.

Table 6. Bank Table

Applicant Name	Card Name	Annual Income
Andrew Smith	ANZ First VISA Card	40,000K
Andrew Smith	Commonwealth Gold Master Card	80,000K

For example in the above table, inconsistency in values. People may show different income while applying for different credit card. Here Andrew Smith has shown different income for applying for two different credit cards of banks. Fuzzy function can be written and membership value will be assigned to each statement to assign the degree of truth.

2.5.2 Inconsistent Data Across Multiple Files/Tables in Database

This type of data arises because integrity constraints that encompass all semantically related tables are not specified and enforced.

2.5.3 Inconsistency Due to Database Overlap

In any environment of multiple databases it is practically unavoidable that the databases would overlap. In multiple databases if one database have one or more elements in common with another database than we can say that databases are overlapped.

2.6 Incomplete Data

Incompleteness usually denotes absence of a value, yet could be regarded as imprecise information where the set of possible values encompasses the entire domain of legal values. Incomplete processing of data occurs when the observation can't perform its function on time [9]. Incomplete Data occurs due to *dirty read, lost update, Unrepeatable Read* and *Missing Values* [11].

3 Conclusion

In this paper we have shown importance of fuzzy logic in database system. We present general-purpose taxonomy of fuzzy data. Fuzzy data representation reflects how, where and to what extent fuzziness is incorporated into classical models. The taxonomy provides framework for understanding the origins of fuzzy data and the impact of fuzzy data in database management system. If you know the possibility that certain fuzzy data exist, you will be more prone to spot them and to plan your project to store & manipulate fuzzy data in a manageable way. We expect that such taxonomy will provide a valuable guideline for further research.

References

1. Altrock, C.: Fuzzy Logic and NeuroFuzzy Applications Explained. Prentice Hall PTR Englewood Cliffs, New Jersey (1995)
2. Buckles, B.P.: An information retrieval perspective on fuzzy database systems. Proceedings of the ACM '82 conference, ISBN:0-89791-085-0, (1982) Pages: 186 – 187
3. Elmasri, R., Navathe, S.: Fundamentals of Database Systems, Addision Wisley, Third Edition, ISBN 9814053309 (2000)
4. Kwan, S., Olken, F., and Rotem, D.: Uncertain, Incomplete and Inconsistent Data in Scinetific and Statistical Databases. Uncertainty Management in Information Systems (1996) 127-154.
5. Mendel, J.M.: Fuzzy logic systems for engineering: a tutorial. Proceedings of the IEEE, Volume: 83 Issue: 3, (1995) pp. 345 –377
6. Parsons, S.: Current approaches to handling imperfect information in data and knowledge bases. Knowledge and Data Engineering IEEE, Vol.8, No. 3, (1996) Page(s): 353-372.
7. Schenker, A., Last, M. and Kandel, A.: Fuzzification of an Object-Oriented Database System. International Journal of Fuzzy Systems, (2001) Vol. 3, No. 2.
8. Shenoi, S.: Multilevel Database Security Using Information Clouding. Second IEEE international conference on Fuzzy Systems, Vol. 1, (1993) pp. 483-488.
9. Smets, P., Motro, A.: Uncertainty Management in Information Systems From Needs to Solutions. Kluwer Academic Publishers (1997)
10. Zaniolo, C.: Database Relations with Null Values. Symposium on Principles of Database Systems Vol.28, No1 (1984).
11. Zhang, W., Wang, K.: An efficient evaluation of a fuzzy equi-join using fuzzy equality indicators. Knowledge and Data Engineering, IEEE Transactions on, Volume: 12 Issue: 2, (2000) Page(s): 225 –237